

A U.S. Decennial Census Browser Shiny application: Visualizing the African American population over time

Kiegan Rice, Iowa State University

Census data provide an important snapshot of information about a country at different times throughout its history, and their value is difficult to overstate. The evolving and adapting nature of census data creates quite the complicated data management problem, and this problem is only exaggerated when one is interested in visualizing data across several years simultaneously. However, improving methods available for data analysis and visualization provide a window for developing a more clean interface for census data exploration. I present a Shiny application that allows users to browse U.S. Census Data across multiple years simultaneously, create an initial visualization of variables of interest, and download chosen data in a ‘tidy’ format. Using an example dataset downloaded from this Shiny app, I then present a more in-depth set of visuals for the African American population between 1790 and 1960 in the United States.

Introduction

Census data provide an important snapshot of information about a country at different times throughout its history, and their value is difficult to overstate. While history books present a narrative of the events that occurred, students themselves often don’t get to interact with the raw data in that learning environment. Clean and accessible census data allow the exploration of different demographic groups over time, or investigations of a particular period of time and what the demographic and economic landscape looked like in the past. Even today, data about the world around us opens a pathway for learning more about places we haven’t been and groups of people with whom we may not usually engage.

From an early point in the United States’ history, there were many “eminent men of science” who recognized the value of the census data and worked to aggregate and present those data. Francis A. Walker’s “Statistical Atlas of the United States”, based on the 1870 census, was an impressive effort in aggregating population data to present them in a visually appealing way [5]. Although the Census Bureau’s “Statistical Atlases” eventually stopped being created, they were an important start to the effort of visually presenting census data to a wider public (see Figure 1).

Today, as methods of data analysis and visualization continue to be developed and improved, access to census data allows us to look back on that history and explore, synthesize, and visualize the information. When aggregated and presented in a clear manner, viewers can learn more about patterns in many different parts of the population. A 2007 paper about the Statistical Atlas discusses just how much information is present in those original

charts and demonstrates several other ways of presenting the information that can be found therein [2].

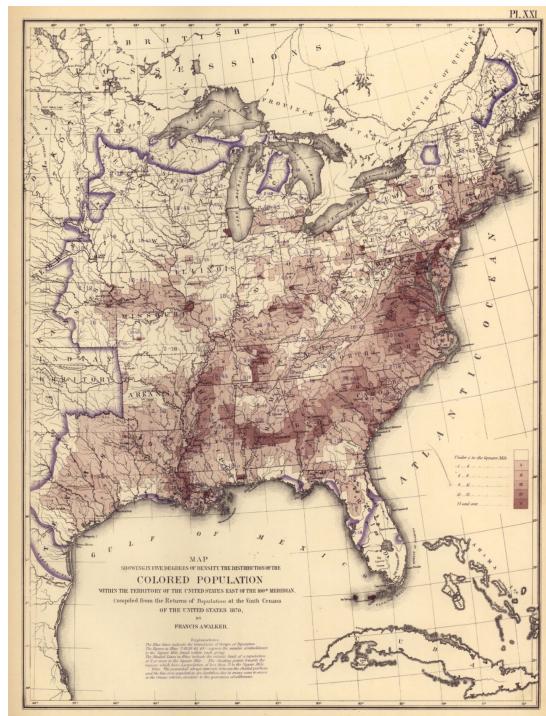


Figure 1: A visualization of the African American population in the United States from the “Statistical Atlas of the United States”.

Ever-improving visualization and data-wrangling methods in R [6] give those interested in statistical graphics a wealth of opportunities to explore and learn from data; in particular, incorporating user interactivity using Shiny has revolutionized the way statisticians share and communicate information [1]. However, it is difficult to make use of these tools with census data if those data are not available and easy to explore in one location.

An inherent problem in census data is that a country’s census changes over time; the variables collected, how they are collected, and even the locations on which they are collected are updated as the country is formed, and subsequently grows and changes. The United States census data are no exception to this rule. In little under two and a half centuries, the census has taken many different forms. Data on occupations have transformed as the employment landscape has changed; new states have been formed, the most recent being within the last 100 years; definitions of various demographic groups and the terminology used to describe them have been updated as the demographic makeup of the country has changed. Each decennial census brings a different set of variables to the table. Sometimes these variables are new things the Census Bureau is interested in learning about, while sometimes they remove variables that are no longer relevant or whose information

is captured elsewhere.

Unfortunately, because the founders of the U.S. Census were unable to see 200 years into the future, those interested in working with census data are left with quite the inescapable mess. If you want to focus in on a particular demographic group and their journey as part of the population of the United States, you may have ten or more different variable names to describe that one group over the course of the census from 1790 to 1960 - and that is just for a single group! Of course, we cannot just simply change variable names to match our own research needs. It is important to keep the data in their true form and be respectful of the way in which the population was defined at different times throughout history, even if our instinct may be to ‘clean’ the data by changing variable names.

This, of course, leaves the user with a wide variety of variables that are far from consistent across years. In order to track one demographic group across years - let alone many groups - a clean user interface that helps users see exactly what information is available to them is a necessity. To streamline the process and assist researchers in finding out to what information they have access and what information they lack, a U.S. Decennial Census Browser is presented as a Shiny application, with the ability to explore available variables within a certain year or across several years. Additionally, a visualization tab allows users to immediately visualize the variables they have picked across different years to get an initial look before downloading the data in the form of ‘tidy’ comma-separated values (csv) files.

I use this application to investigate trends in the African American population over time in the United States. First, this is done within the visualization tools provided by the Shiny application for a small subset of years. I subsequently download the relevant data and take a closer look at trends over the entire range of available data.

Data Access

There are two main datasets that contain aggregated counts of historical, demographic, economic, and social characteristics from the United States decennial census. They were both collected and developed by the Inter-University Consortium for Political and Social Research (ICPSR). The ICPSR 3, gathered from computer-readable data collections from the U.S. Census Bureau as well as other reports (both published and unpublished), contains data from 1790 to 1960 (*do i need to put a citation here?*). The ICPSR 2896 includes much of the same information as the ICPSR 3, and in addition also includes a wider array of variables such as manufacturing and more county and city-level information. The ICPSR 2896 is a restricted-access dataset that requires users be part of a member institution in order to gain access.

The University of Virginia Library hosted a “Historical Census Browser” for many years that allowed users to search United States Decennial Census Data for research, teaching, and personal purposes [4]. The data included records on various aspects of the U.S. population from the 1790 Census through the 1960 Census, originally populated

using the ICPSR 3 dataset. This Historical Census Browser was free and available for use by anyone with an internet connection. The browser allowed a user to peruse available topics for each census year at both the state and county levels. The Historical Census Browser was taken down on December 31, 2016, but the county-level aggregated data had already become unavailable several months before then.

Although the county-level data had already been removed from the website, we captured the state-level aggregated records for each decennial census from 1790 through 1960 from the website in September of 2016 and saved them as raw data. These data are used to populate the Shiny application.

The Census Browser Shiny application

The Shiny application has three main points of focus, split into separate tabs. The first tab, ‘Single Year’, allows users to explore all available variable names for a single year of the decennial census at a time. After choosing a particular year, users can look for up to two specific variables of interest by using two available text search bars. This will narrow down the possible variable names, allowing for a more narrow focus in a given year.

The second tab, for ‘Multiple Years’, brings in the ability to compare availability of variables across several years. Users may select a range of years and, if they so choose, narrow down the given results with a text search term. These results are presented as a table with all chosen years as ‘columns’, while all relevant variable names appear as ‘rows’ in the table (see Figure 2). An X indicates the presence of that particular variable name in the corresponding year. A totals column to the far right allows users to not only see in how many years each variable appears, but also order the relevant variables by that frequency.

Available_Colnames	1800	1810	1820	1830	Total
SLAVES	X	X			2
TOTAL.SLAVES		X	X		2
FEMALE.SLAVES.26-44.YEARS.OF.AGE		X			1
FEMALE.SLAVES.14-25.YEARS.OF.AGE		X			1
FEMALE.SLAVES.UNDER.14.YEARS.OF.AGE		X			1
MALE.SLAVES.45.YEARS.OF.AGE AND OVER		X			1
MALE.SLAVES.26-44.YEARS.OF.AGE		X			1
MALE.SLAVES.14-25.YEARS.OF.AGE		X			1
MALE.SLAVES.UNDER.14.YEARS.OF.AGE		X			1
FEMALE.SLAVES		X			1

Figure 2: Resulting view of the ‘Multiple Years’ tab after restricting the range of years to between 1800 and 1830, and searching for the term ‘slave’.

The final tab, ‘Visualize’, creates an on-the-spot visual representation of the data selected in the ‘Multiple Years’ portion of the application. Users may choose to simply separate data by year, and sum over all variables chosen

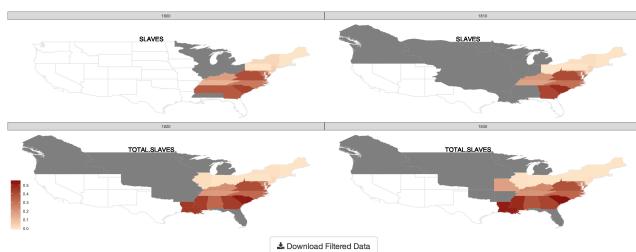


Figure 3: A visual representation of chosen variables for total number of slaves between the years 1800 and 1830, with period-accurate mapping for state boundaries. Maps are labeled with the variable name used. Values are shown as a percentage of total population in the state at that time.

within each year (see Figure 3), or facet by both year and variable chosen (see Figure 4). Depending on the range of years selected by the user and the number of different variables chosen, the latter can easily become a very large grid environment. However, it still allows users to see several things about the particular variables and years they are looking at .

The **USAboundaries** package was used to create period-accurate representations for each separate year [3]. First of all, every plot has the outlines of the 48 contiguous states as of 1999. Above that base, the outlines for each state are plotted as they were in the year represented. This means that users can see varying state boundaries across years - however, since these are the areas data were actually collected on at the time, they are the true representation of the geographic regions for the demographic groups chosen. In some years, there are states that still hold Territory status, or are not yet included in the decennial census. If users have chosen to sum over chosen variables for each year, these will show up with the other states, but be shown in grey as they contain **NA** values. If users have chosen to look at the expanded grid, these will show up in the bottom row. Both situations provide a useful look at where those boundaries existed. Users can immediately see how they change over time and see demographic groups represented correctly in the areas they were collected on.

Secondly, users can see how variable names change over time. They can see this during the initial selection of variables in the ‘Multiple Years’ tab; however, now they can see differing variable names for an equivalent or similar demographic group and still directly compare the values on the maps.

While this representation may not be without faults in some scenarios, and can easily become an overwhelming graphic with many different variables represented, we can get an initial visualization of our selected variables with very minimal effort. Choosing to download the filtered data and work with it ourselves in R proves even more fruitful.

For the example of African Americans throughout U.S. history, we need several different search terms. Three searches were executed within the data, for variables containing the terms **SLAVE**, **COLORED**, and **NEGRO**. For each of these, all variables that counted total numbers within

these categories were selected, as well as some that split by gender if the ‘total’ variable was not available for that year.

After downloading each of these sets of selected variables, they were read in to R and combined using the **full_join** function from the **dplyr** package *Should I cite dplyr?*. Downloaded data are kept in a ‘tidy’ fashion and each file automatically includes the **YEAR**, **STATE**, **TOTAL.POPULATION**, and **TYPE** variables, the last of which indicates whether that state had yet earned statehood or was still considered a territory. This makes the files an excellent candidate for **full_join**, since each row contains a year-state pair. In this case, this is advantageous since there are years in which there are variables for both the **SLAVE** and **COLORED** terms, and they can now reside together in one common row even though they originated from separate searches and were downloaded as separate csv files.

With this combined data set, we now have a wealth of information to use as a starting point for our exploration of the African American population throughout U.S. history. We can also create more detailed graphics and create comparisons between different variables.

A Closer Look at the Data

Within our combined data set, we can filter our data on each year as we look through it to be able to quickly see what variables hold values for that particular year. In the early years of the U.S. decennial census, we are almost exclusively left with variables with the term **SLAVE** in them. The vast majority of African Americans were slaves when the United States was founded, so variables that count the number of slaves tell the story about the African American population in early U.S. history. Note that since the **SLAVE** categorization is also the only term the U.S. census had related to African Americans in the early years of the census, it is the only source of information available on that group for that time period.

Beginning in the year 1790, we see the term **SLAVES** is our main source of information. Using a similar tactic to that portrayed in the ‘Visualize’ tab of the Shiny application, we can plot the current state outlines as a base, followed by 1790 values for the **SLAVES** variable in another layer, with 1790-specific state outlines. The resulting image, seen in Figure 5, provides a basic look at a single variable in a single year. Moving forward, the ways in which the African American population is accounted for in the census gets more complicated; thus, we will need to consider additional ways of exploring the information.

For example, let us examine the year 1820. In the first half of the 19th century, slavery was still quite prevalent in southern states. However, there were also some free colored persons residing in most states at that time. Thus, we can now expect to see at least two separate categories to capture this particular demographic. At this time in U.S. history, most non-white minorities were grouped together in the “colored persons” category, and therefore we do not have any other record for African Americans except the grouped “colored persons” categorization and the “slaves”

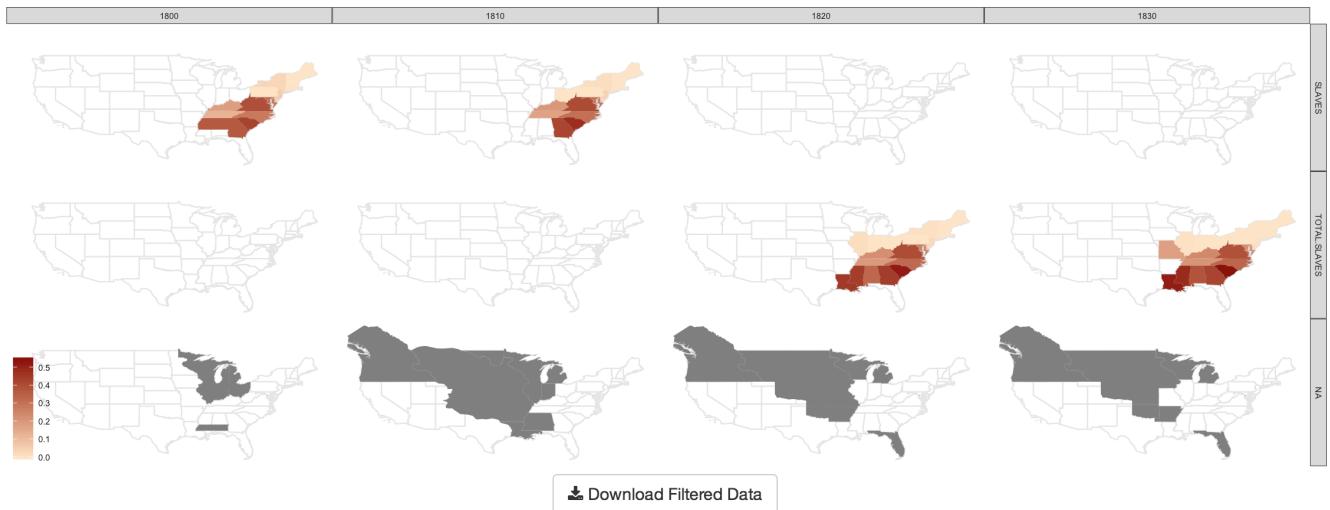


Figure 4: A visual representation of chosen variables for total number of slaves between the years 1800 and 1830, with period-accurate mapping for state boundaries. Maps are faceted by variable name used to create the map. Values are shown as a percentage of total population in the state at that time.

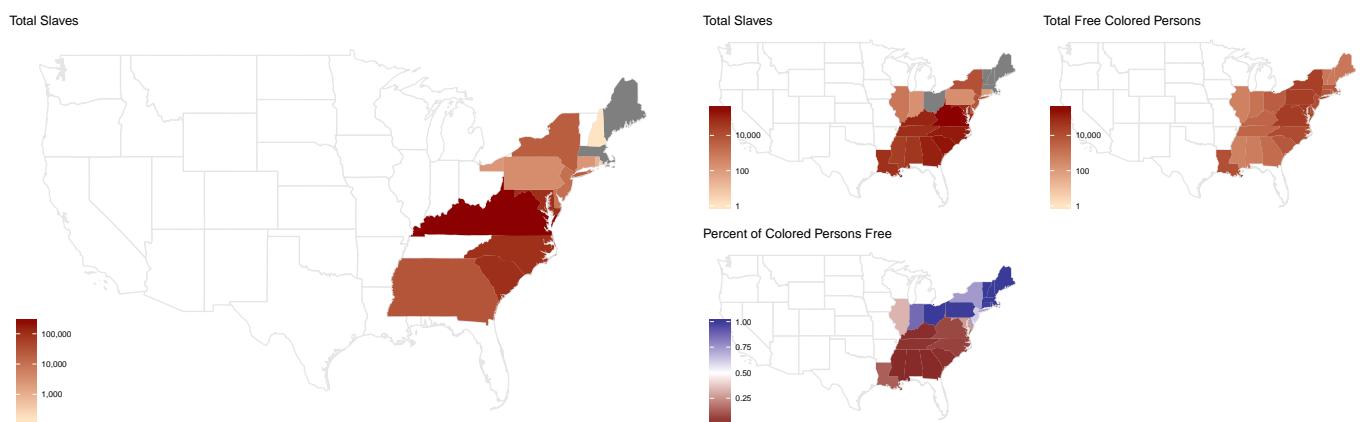


Figure 5: Total number of slaves per state in 1790, plotted on a continuous log scale. State boundaries for July 4, 1790 were gathered from ‘USAboundaries’ package.

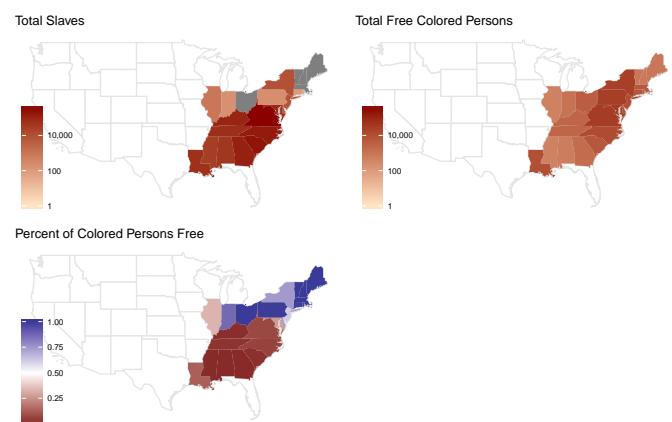


Figure 6: Total number of slaves, total number of free colored persons, and percentage of colored persons free in each state in 1820 United States.

categorization.

Upon examining the available variables, we observe a count of **TOTAL.FREE.COLORED.PERSONS** as well as a count of **TOTAL.SLAVES**. This gives us the ability to visualize not only the total numbers in each of these categories within each state, but also the percentage of colored persons in each state that were categorized as “free persons” as opposed to slaves (see Figure 6). Knowing that both of these variables are available allows for a powerful comparison by visualizing the growing divide between northern and southern states in the United States at that time. Although the census recorded values of **TOTAL.FREE.COLORED.PERSONS** in each state, there was still a huge majority enslaved in most southern states at this time.

Moving forward in time, even more information can be gleaned about this population. When we subset our data on the year 1850, we see that although there is still a **TOTAL.SLAVES** column in the record, many states no longer record this variable; by then, they had already abolished slavery. Because this is a specific situation in which we

know why there are ‘missing’ data values, we can transform these NA values into zeros to account for this difference in the data and allow us to still calculate the percentages of free colored persons. Additionally, since free colored persons are becoming more prevalent in the data, we can also start investigating what the total colored population looks like in each state, beyond just the size of the slave population and freed population (see Figure 7).

Due to the abolition of slavery at the national level in 1865, it is also valuable to explore how the overall population of colored persons changes in each state throughout several years in the 19th century. Some migration begins to occur out of some southern states. Although the highest densities of the population of colored persons remains in southeastern states, a shift can be observed towards some northern states. A separate shift can also be seen out towards California, during a period of major gold mining operations. It is interesting to observe this overall shift of the total population of colored persons in each state as the United States transitions into the post-slavery period

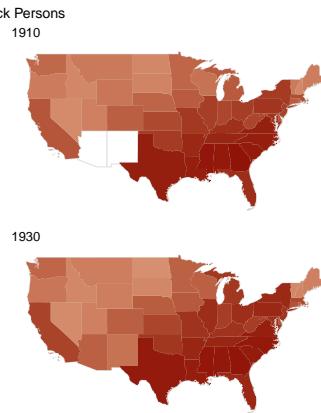
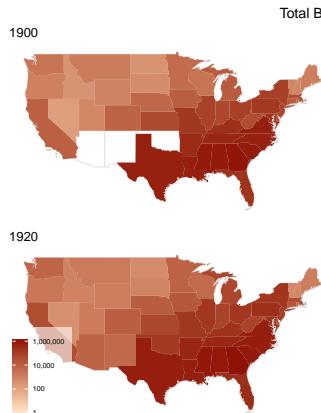
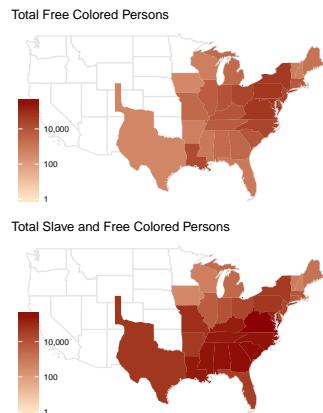
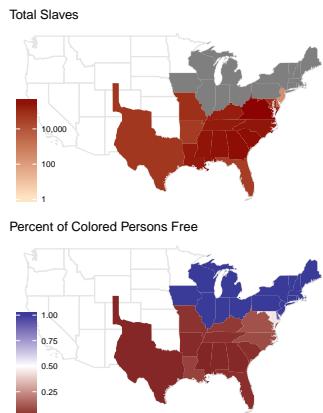


Figure 7: Total number of slaves, total number of free colored persons, and percentage of colored persons free in each state in 1850 United States.

(see Figure 8). We also see how the boundaries of states change and new states are formed, with the United States gaining new states at each census.

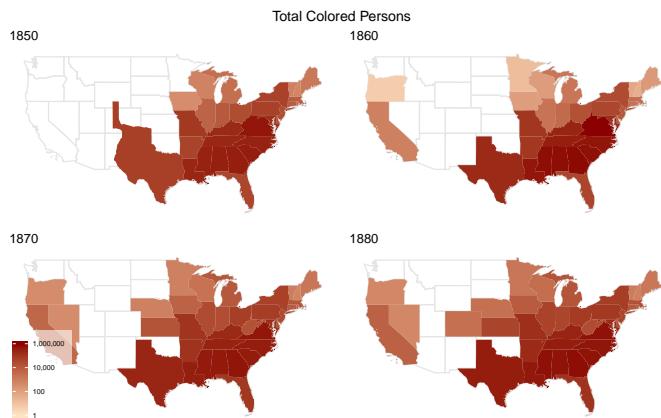


Figure 8: Total number of colored persons per state in 19th Century United States.

We end our investigation with a look at the beginning of the 20th century. For the first time in the census data, we see the arrival of variables referring specifically to the African American population. These variables are denoted in the original data with terminology such as **TOTAL.NEGROES**. In visualizations, we will use these variables to represent ‘total black persons’; this is a clear example of changing terminology and the need to navigate census data with caution.

In this final series of visuals, we can see the continuing shift of the African American population (which we previously observed as one part of the ‘colored’ population) out west to California, and continuing northward movement in the east and midwest (see Figure 9). Although the highest density of African American persons remains in the southern states, we see a significant northern migration taking place.

Conclusion

This section still needs some work.

Figure 9: Total number of black persons per state in 20th Century United States.

Even when tracking one particular group, there is a huge lack of consistency within the data. For African Americans in the United States, there isn’t even a single specific category to represent them until the beginning of the 20th century. Before that, tracking the African American population is very difficult without having some context for how terminologies and definition of demographic groups has changed. And even once a specific group is created within the census, there are continuing changes with variable naming; some years include a total count while some include a count of males and females, but no total. And definitions of demographic groups and the manner in which they are recorded continue to change to this day.

The decennial census is a rich source of historical information. We have built a tool that can help extract and visualize that information to, for example, understand how demographic attributes of the U.S. population have changed over time.

Although there is no county-level information available and the data for the census browser go only through 1960, users are able to efficiently explore variables of interest at different time points in U.S. history and tell a visual story about a growing nation and a changing population. This connectivity across years and ability to assess what information was available when, all in one browser, have the potential to save time and effort of researchers interested in delving into United States history.

The structure of our browsers also helps determine not only the variables that are available in a given year, but also where information is missing for some states on a particular variable. Finding these differences is easier when all relevant variables are combined in one data frame, which can be summarized and searched all at one time. These differences are even more clear in the ‘visualize’ tab, where users can see not only which variables are relevant for different years, but also see whether states have missing data in the chosen years.

References

- [1] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. shiny: Web application framework for r., 2017.

- [2] H. Hofmann. Interview with a centennial chart.
CHANCE, 20:26–35, 2007.
- [3] L. Mullen and J. Bratt. Usaboundaries: Historical and contemporary boundaries of the united states of america, 2017.
- [4] U. of Virginia Library. Historical census browser, 2017.
- [5] U. S. C. Office and F. A. Walker. *Statistical atlas of the United States based on the results of the ninth census 1870 with contributions from many eminent men of science and several departments of the government*. 1874.
- [6] R Core Team. R: A language and environment for statistical computing. 2015.

Further Reading